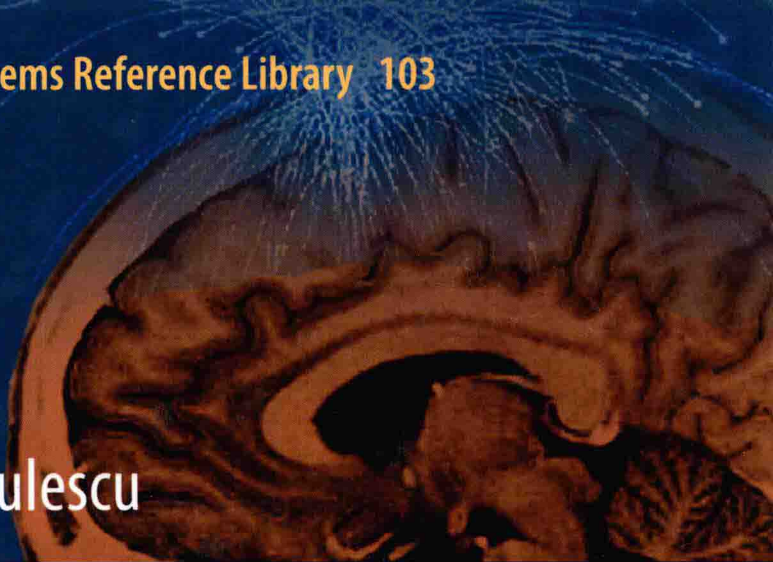
Intelligent Systems Reference Library 103



Alina Bărbulescu

Studies on Time Series Applications in Environmental Sciences

Springer

Alina Bărbulescu

Studies on Time Series Applications in Environmental Sciences

Springer

Alina Bărbulescu
Ovidius University of Constanta
Constanta
Romania

ISSN 1868-4394 ISSN 1868-4408 (electronic)
Intelligent Systems Reference Library
ISBN 978-3-319-30434-2 ISBN 978-3-319-30436-6 (eBook)
DOI 10.1007/978-3-319-30436-6

Library of Congress Control Number: 2016933239

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Intelligent Systems Reference Library

Volume 103

Series editors

About this Series

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included.

More information about this series at http://www.springer.com/series/8578

To Prof. Dr. Eng. Radu Dobrot—
To ***the One***,
that made me laugh in the most difficult
moments,
that encouraged and supported me—
with all my gratitude and love

Preface

Modeling and forecasting hydro-meteorological time series are of great interest due to their practical applications and impact on the human life. Many methods have been successfully used for solving different types of problems in this area. Here, we try only to summarize some possible approaches and to present a part of our results concerning modeling particular hydrological time series, from a region less studied in Europe.

The book is divided into seven chapters. In the first one we introduce the data series. The next two chapters are mainly theoretical. They contain some tests used for checking different statistical hypotheses on univariate time series and their implementation in R software, as well as a very short overview on the modeling techniques applied in the next chapters.

The rest of this book contains a part of our results, obtained the last 7 years on modeling the precipitation series or applications of some methods proposed by other scientists for generation precipitation fields.

One chapter summarizes the results of modeling the pollutants' dissipation in the atmosphere, while another one summarizes that of the evolution of water quality of two lakes, which are known to be affected by the atmospheric pollution and human activities.

All models refer to series from Dobrogea, a region situated in the southeastern part of Romania, between the Black Sea and the Danube River, for which no systematic study of climatic evolution has been done till 2007.

Scientific research is usually a team work, so a part of the results presented here has been obtained in cooperation with Dr. Lucica Barbeş, Dr. Elena Băutu, Dr. Judicael Deguenon, Dr. Cristina Gherghina, Dr. Carmen Elena Maftei, Dr. Elena Pelican, Dr. Nicolae Popescu-Bodorin, and Dr. Dana Simian. Thanks for their cooperation.

All the gratitude goes to Prof. Dr. Eng. Radu Victor Drobot, my Ph.D. supervisor in Civil Engineering. Without his valuable suggestions on different research topics and his continuous encouragement, this book wouldn't exist.

Thanks to Dr. Lakhmi Jain for his invitation to write this book.

Last but not least, thanks to my family that supported me unconditionally.

Sharjah, UAE
October 2015

Alina Bărbulescu

Data Series

The study region is Dobrogea, which is situated in the southeastern part of Romania, between the Black Sea and the Danube River, between 27°15′05″ and 29°30′10″ Eastern longitude and 43°40′04″ and 45°25′03″ Northern latitude. It has a surface of 11,145 km^2 without the Danube Delta and the lake basin Razim-Sinoe [7] (Fig. 1).

Fig. 1 Dobrogea region, and the hydrological and hydro-meteorological stations. The main stations have roman letters attached

Its active surface is not uniform and it presents interesting geo-morphological characteristics. Since the presentation of these features is out the scope of this book, we invite the interested readers to refer to [4, 7].

The average annual temperature is 11 °C in the western part of the territory and is over 11 °C in its northeastern and northern parts. The mean annual temperature decreases from south to north, concomitantly with the altitude increasing and the augmentation of the continental influence inside region [5].

Different aspects related to the temperature variations and modeling of climate in this region can be found in [1–3, 6].

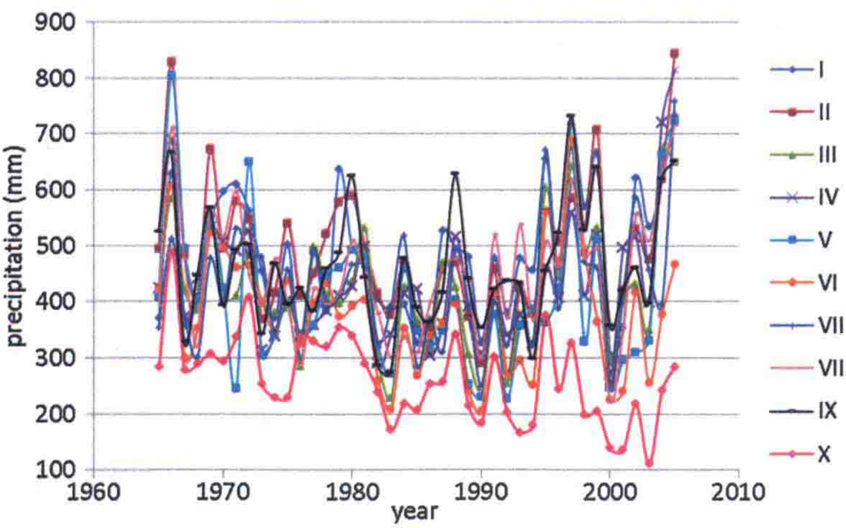


Fig. 2 Annual precipitation series recorded at the main stations

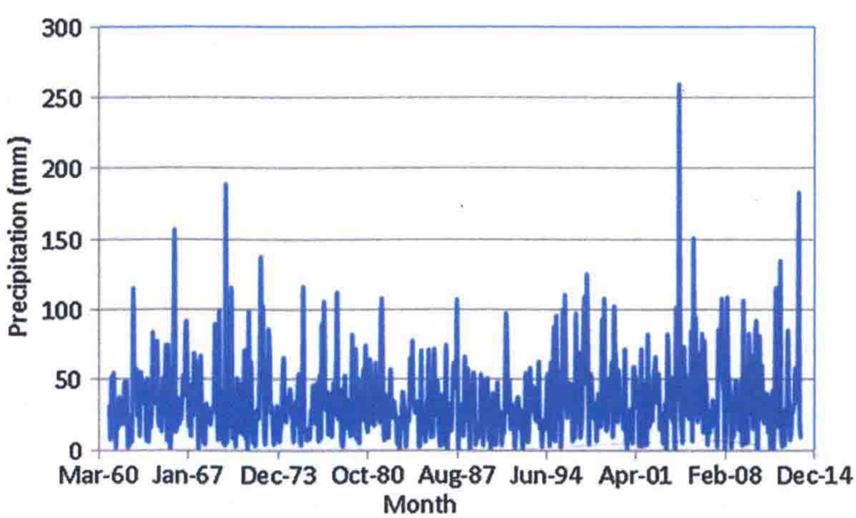


Fig. 3 Constanta monthly precipitation series (January 1961–December 2013)

The annual and monthly series studied here are recorded in the period January 1965–December 2005 at the ten main stations (that has a roman letter attached on the map on Fig. 1) and 41 secondary ones. In Fig. 2, we present the annual series of precipitation record in the period 1995–2005. For some series, as Constanta (Fig. 3) and Sulina, longer periods are used. All the series are complete, without gaps and reliable, collected by INHGA or from http://eca.knmi.nl/.

References

1. Bărbulescu, A.: Modeling temperature evolution. Case study, Romanian Reports in Physics, **68**(1), (2016) (to appear)
2. Bărbulescu, A., Băutu, E.: ARIMA and GEP models for climate variation, International Journal of Mathematics and Computation, **3**(J09), pp. 1–7 (2009)
3. Bărbulescu, A., Băutu, E.: Mathematical models of climate evolution in Dobrudja, Theoretical and Applied Climatology, **100**(1–2), pp. 29–44 (2010)
4. Bărbulescu, A., Deguenon, J., Teodorescu, D.: Study on Water Resources in the Black Sea Region, Nova Publishers, USA (2011)
5. Bărbulescu, A., Teodorescu, D., Deguenon, J.: Study on water resources in the Romanian Black Sea region. In Ryann, A.M., Perkins, N.J. (eds), The Black Sea: Dynamics, Ecology and Conservation, pp. 175–205. Nova Publishers, USA (2011)
6. Maftei, C., Bărbulescu, A.: Statistical analysis of climate evolution in Dobrudja region, World Congress on Engineering, Imperial College of London, England, July 02–04, 2008, Book series: Lecture Notes in Engineering and Computer sciences, vol. II, pp. 1082–1087 (2008)
7. Posea, G., Bogdan,O., Zăvoianu, I.: Geografia Romaniei, vol. **V**, Editura Academiei Române, Bucuresti, (2005) (in Romanian)

Contents

Chapter 1
Hypotheses Testing on Meteorological Time Series

Data analysis is an important step in time series modeling and forecasting. It requires: obtaining and preparing the dataset, exploratory analysis of data series, performing statistical tests and the results' interpretation, the former being of major importance for all the other stages of analysis. Dealing with hydro-meteorological time series requires attention to data acquisition (measurement methodology and frequency), the series length (at least 50 years, in the case of studies concerning the climate change) and their completeness [105].

In this chapter, we focus on testing some statistical hypothesis and methods of detecting the long range dependence property in hydro-meteorological time series.

In what follows we denote the observed data by $(x_i)_{i=\overline{1,n}}$. They are realizations of a time series process, denoted by $(X_i)_{i=\overline{1,n}}$.

1 Normality Tests

The Gaussian distribution is well-known, its properties being established for many years ago. It is a reference distribution to which one may report the experimental data to discover their properties. Testing the series normality is also important because many statistical methods rely on the hypothesis that the series are Gaussian.

The simplest way to check the null hypothesis (H_0: the series is normally distributed) against its alternative (H_1: the series is not normally distributed) is the use of graphical representations. One of them is the quantile–quantile plot (Q-Q plot) employed for deciding whether a univariate random sample comes from a given distribution G. The Q-Q plot is obtained by plotting the quantiles of the sample against the theoretical quantiles of G. If the sample comes from the specified distribution (in our case, the Gaussian one) then the points are close to a straight line.

MINITAB, SPSS, R have the option to draw the Q-Q plot. We mention that R is freeware software.

A. Bărbulescu, *Studies on Time Series Applications in Environmental Sciences*,
Intelligent Systems Reference Library 103, DOI 10.1007/978-3-319-30436-6_1

In the following we present the R code for obtaining the Q-Q plot of Constanta annual precipitation series (1961–2013).

```
data<- read.csv("D:\\Lucrari_2.12.14\\2015_Carte\\Cta_annual_1961_2013.csv", sep=",",
header=TRUE)
qqnorm(y); qqline(y, col = 2)
```

Looking to the chart (Fig. 1a) we could not reject H_0. We couldn't also decide the opposite since the points from the upper part of the Q-Q plot are not very close to the line.

The same decision could be taken, looking to the histogram of the series (Fig. 1b), which is a two-dimensional representation of the observed data against their frequency.

The R-function used to create the histogram is: `hist(y, right=FALSE)`.

The decision to reject the normality hypothesis can be taken very easy for Constanta monthly series (1961–2013) because the plots significantly deviate from the straight line and the histogram is right-skewed (Fig. 2).

Since the decision on data normality is still difficult for Constanta annual series, other methods, based on statistical tests, are more appropriate. Among the tests used for this aim we mention: Kolmogorov–Smirnov [78, 111], Jarque–Bera [72], Shapiro–Wilk [110], Lilliefors [82] (presented in detail in [7]), Anderson–Darling [2] and Cramer von Mises [113].

The Anderson–Darling test is used to determine if a dataset comes from a specified distribution (the normal one, in our case), so it is considered to be also a goodness of fit test. It compares the fit of an observed cumulative distribution function to the expected cumulative distribution function.

Considering the sample data $\{x_i\}_{i=\overline{1,n}}$, and $\{x_{(i)}\}_{i=\overline{1,n}}$ its values increasingly ordered, the Anderson–Darling statistic is defined by:

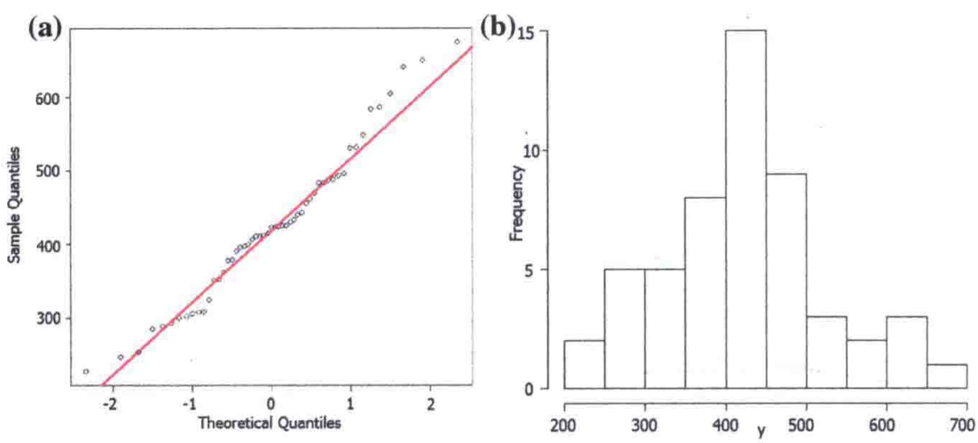


Fig. 1 **a** Q-Q plot and **b** histogram of Constanta annual precipitation series (1961–2013)

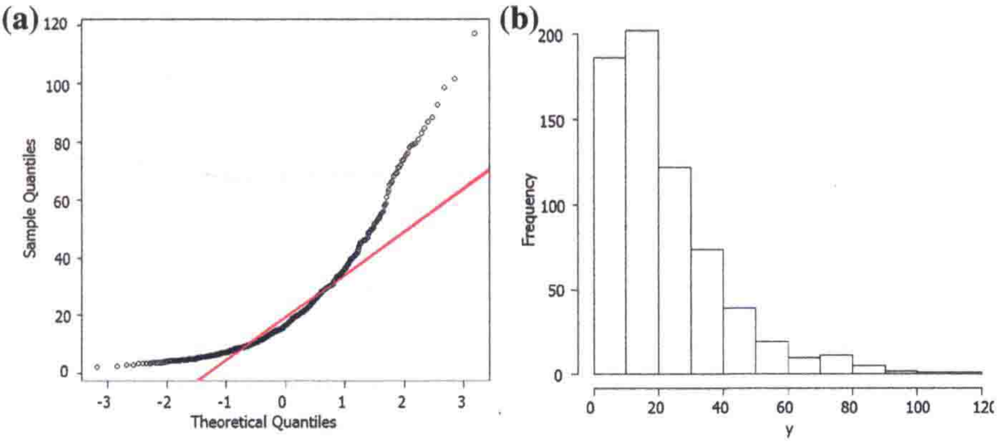


Fig. 2 **a** Q-Q plot and **b** histogram of Constanta monthly precipitation series (1961–2013)

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(x_{(i)}) + \ln(1 - F(x_{(n-i+1)}))], \tag{1}$$

where F is the cumulative distribution function of the specified distribution.

If Anderson–Darling is used as normality test, (1) becomes:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln p_{(i)} + \ln(1 - p_{(n-i+1)})],$$

where:

$$p_{(i)} = \Phi([x_{(i)} - \bar{x}]/s), \tag{2}$$

Φ being the cumulative distribution function of the standard normal distribution, $\bar{x}$, the mean and s—the standard deviation of $\{x_i\}_{i=\overline{1,n}}$.

The Cramer-von Mises test for the composite hypothesis of normality is also based on the cumulative distribution function. The test statistic is:

$$W = \frac{1}{12n} + \sum_{i=1}^{n}\left(p_{(i)} - \frac{2i-1}{2n}\right),$$

where $p_{(i)}$ is given in (2).

The Pearson chi-square test for normality is applied to binned data so that the value of the statistic of the test depends on how the data was binned. Firstly, the data are standardized by subtracting the sample mean and dividing each value by the sample standard deviation. Then, the number of bins (k) is chosen by a formula (there is no optimal formula for this choice!), as, for example, $k = 1 + \log_2 n$, where n is the sample data.

The test statistic is:

$$\chi^2 = \sum_{i=1}^{k} (o_i - e_i)^2 / e_i,$$

where o_i is the observed frequency of the ith bin, e_i is the expected frequency of the ith bin, calculated by:

$$e_i = F(x_2) - F(x_1),$$

F is the cumulative distribution function of the distribution being tested, and x_1, x_2 are the limits of the ith bin.

When the mean and variance are known, the test statistic is asymptotically χ^2 distributed with $k - 1$ degrees of freedom. Therefore, the null hypothesis is rejected at a significance level α if the test statistic is greater than the quartile $\chi^2_{1-\alpha}(k-1)$.

Simulation studies showed that the normality tests have different powers. The relative powers of the discrete statistics tests of Kolmogorov–Smirnov, Cramér-von Mises, Anderson–Darling and Watson and of two test statistics for nominal data (chi-square and Kolmogorov–Smirnov) for an uniform null distribution against a selection of fully specified alternative distributions has been studied in [112]. The results show that the Pearson's chi-square [96] and the nominal Kolmogorov–Smirnov are more powerful for the studied triangular, sharp and bimodal alternative distributions.

In the same idea, to compare the power of the Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests, the Monte Carlo procedure was employed in [102]. The results proved that the least powerful test is the Kolmogorov–Smirnov one, and the most powerful one, Shapiro–Wilk, followed by Anderson–Darling. However, the Shapiro–Wilk test is biased if the sample size and the significance level are small (that is the power could be less than α). Due to its inferior power by comparison to the other tests, it is not advisable to use the Pearson chi-square test for checking the composite hypothesis of normality [62, 91].

In the following, the statistical tests are performed at a significance level of 0.05, if another level is not specified.

To perform normality tests using the R software, one has to install the packages **fBasics** [55] and **nortest** [62] and to write the following commands: